Pattern Anal Applic (2013) 16:99–116
DOI 10.1007/s10044-011-0199-9

THEORETICAL ADVANCES

On utilizing dependence-based information to enhance micro-aggregation for secure statistical databases

B. John Oommen · Ebaa Fayyoumi

Received: 2 May 2010 / Accepted: 16 February 2011 / Published online: 16 March 2011

Abstract We consider the micro-aggregation problem which involves partitioning a set of individual records in a micro-data file into a number of mutually exclusive and exhaustive groups. This problem, which seeks for the best partition of the micro-data file, is known to be NP-hard, and has been tackled using many heuristic solutions. In this paper, we would like to demonstrate that in the process of developing micro-aggregation techniques (MATs), it is expedient to incorporate information about the dependence between the random variables in the micro-data file. This can be achieved by pre-processing the micro-data *before* invoking any MAT, in order to extract the useful dependence information from the joint probability distribution of the variables in the micro-data file, and then accomplishing the micro-aggregation on the "maximally independent" variables—thus confirming the conjecture [A conjecture, which was recently proposed by Domingo-Ferrer et al. (IEEE Trans Knowl Data Eng 14(1):189–201, 2002), was that the phenomenon of micro-aggregation can be enhanced by incorporating dependence-based information between the random variables of the micro-data file by working with (i.e., selecting) the maximally independent variables. Domingo-Ferrer et al. have proposed to select one variable from among the set of highly correlated variables inferred via the correlation matrix of the micro-data file. In this paper, we demonstrate that this process can be automated, and that it is advantageous to select the "most independent variables" by using methods distinct from those involving the correlation matrix.] of Domingo-Ferrer et al. Our results, on real life and artificial data sets, show that including such information will enhance the process of determining how many variables are to be used, and which of them should be used in the micro-aggregation process.

Keywords Micro-aggregation technique · Maximum spanning tree · Projected variables

1 Introduction

Central to the study of secure statistical databases are a family of algorithms classified in the literature as being "micro-aggregation" techniques (MATs). Apart from being fast and efficient, they are also intuitively appealing because they are akin to the family of clustering methods. This paper considers how such methods can be enhanced, both with regard to "accuracy" and efficiency, by learning, and thereafter incorporating the information that relates to the dependence between the random variables being analyzed. In all brevity, we are not aware of any other reported method which specifically incorporates such dependence-type information to optimize an MAT, or for that matter, to optimize a method which controls the information loss (IL) and the disclosure risk (DR) in secure statistical databases.

A preliminary version of some of the results from this paper appeared in the Proceedings of ACISP'08, the Thirteenth Australasian Conference on Information Security and Privacy, in Wollongong, Australia, in July 2008.

B. J. Oommen (✉) · E. Fayyoumi
School of Computer Science, Carleton University, Ottawa K1S 5B6, Canada
e-mail: oommen@scs.carleton.ca

E. Fayyoumi
e-mail: efayyoum@scs.carleton.ca

B. J. Oommen
University of Agder, Grimstad, Norway

A lot of attention has recently been dedicated to the problem of maintaining the confidentiality of statistical databases through the application of statistical tools, so as to limit the identification of information on individuals and enterprises. Statistical disclosure control (SDC) seeks a balance between the confidentiality and the data utility criteria. For example, federal agencies and their contractors who release statistical tables or micro-data files are often required by law or by established policies to protect the confidentiality of released information. However, this restriction should not affect public policy decisions which are made by accessing only non-confidential summary statistics [1, 20]. Therefore, optimizing the IL and the DR so as to reach an equilibrium point between them is not an easy task[1] [1].

The micro-aggregation problem (MAP), as formulated in [4, 10, 18, 21, 25], can be stated as follows: a micro-data set $\mathcal{U} = \{U_1, U_2, \ldots, U_n\}$ is specified in terms of the n "micro-records", namely the $U_i's$, each representing a data vector whose components are d continuous variables. Each data vector can be viewed as $U_i = [u_{i1}, u_{i2}, \ldots, u_{id}]^{\mathrm{T}}$, where u_{ij} specifies the value of the jth variable in the ith data vector. Micro-aggregation involves partitioning the n data vectors into, say m, mutually exclusive and exhaustive groups so as to obtain a k-partition $\mathbb{P}_k = \{G_i | 1 \leq i \leq m\}$, such that each group, G_i, of size, n_i, contains either k data vectors (fixed-size case) or between k and $2k-1$ data vectors (data-oriented case).

The optimal k-partition, $\mathbb{P}_k^*$, is defined to be the one that maximizes the within-group similarity, which is defined as the *sum of squares error*, $\mathrm{SSE} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^{\mathrm{T}} (X_{ij} - \bar{X}_i)$. This quantity is computed on the basis of the Euclidean distance of each data vector X_{ij} to the centroid $\bar{X}_i$ of the group to which it belongs. The *information loss* is measured as $\mathrm{IL} = \frac{\mathrm{SSE}}{\mathrm{SST}}$, where SST is the squared error that would result if all records were included in a single group, and is given as $\mathrm{SST} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^{\mathrm{T}} (X_{ij} - \bar{X})$, where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. In the literature, the quantity IL is also conveniently specified as a percentage.

Understanding the presence and structure of dependency between a set of random variables is a fundamental problem in the design and analysis of many types of systems including filtering, pattern recognition, etc. As far as we know its application in SDC has been minimal. Utilizing this information is the goal of this paper. Typically, in modern day systems, the data protector has been able to choose the technique and set its parameters without a thorough understanding of the characteristics of the micro-data file, and the stochastic dependence of the variables. Although gleaning this information could be particularly difficult and even time-consuming, our hypothesis is that this information is central to the micro-data file, especially when working in a high dimensional space.

Undoubtedly, the IL is minimized when all the variables are included in the MAT. Otherwise, the result of the multivariate MAT depends on the number of variables used in the micro-aggregation process. However, more recent research (see for example [10, 13]) have recommended studying the *dependence* between the variables themselves. Indeed, the prior art has been reported that the multivariate micro-aggregation on un-projected data taking two, three or four variables offers the best trade-off between the IL and the DR (i.e., within the limited setting of not incorporating the information in all the variables). In other words, deciding on the number of variables to be taken into account, and on the *identity* of the variables to be micro-aggregated, is far from trivial. Domingo-Ferrer and Torra [13] have reported that multi-variate micro-aggregation on unprojected data taking two or three variables at a time (rather than incorporating the information in all the variables) offers the best trade-off between IL and DR. The unanswered question is that of inferring which variables should be used in this process. Indeed, we believe that a solution to this puzzle lies in the inter-variable "dependence" information, as confirmed by the works of Nin et al. [26].

Sanchez et al. [30] have emphasized that the decision about which variables are to be chosen has to be gleaned from a priori "knowledge about the characteristics of each variable from the experts". While this is a feasible approach, we argue that it is subjective, and that a formal objective method is desirable. Indeed, what will happen if the researcher encounters a new project for which there is no prior knowledge? Or how we will proceed if an expert for a specific data domain is not available? Our aim is to minimize the necessity to depend on a human expert, but rather to have the ability to study and estimate the characteristics of each variable objectively. Thus, we seek a systematic process by which we can choose the desired variables automatically and, thereafter, micro-aggregate the file.

This paper involves MATs, but rather from a perspective different than the ones that have been considered in the literature. We propose a scheme by which we can avoid using the information in *all* the dimensions (for example, in computing the distance between 2 records, etc.). Furthermore, neither will we resort to projecting the micro-data file onto a single axis, nor will we attempt to micro-aggregate it using any specific sorting method [6–9, 23–25, 29]. The main contribution of this paper is to extract useful information from the joint probability distribution of the variables in the file to be micro-aggregated. Then, rather

[1] The review presented here has been abridged as per the advice of the Referees.

than use *all* the variables in the micro-data file, we propose to only process the "maximally independent" variables in the subsequent multi-variate micro-aggregation. Indeed, we propose to use such a method as a pre-processing step before *any* MAT is invoked, and to test the effect of using such a dependency analysis on the micro-aggregation process so as to reduce the computational time, and IL.[2]

The structure of this paper is as follows: in Sect. 2 we summarize the background about the most recent MATs and, in particular, the maximum distance to average vector (MDAV) scheme. In Sect. 3 the enhanced micro-aggregation dependence is presented informally and algorithmically. Then, in Sect. 4, we present the results of experiments we have carried out for synthetic and real data sets. The paper finishes in Sect. 5 with some conclusions.

2 Background

In this section, we start with a brief but concise survey[3] about the reported MATs. Subsequently, we present a brief description of the MDAV method, which will be used after invoking the pre-processing step which specifies the number and the identity of each variable to be used in the MDAV micro-aggregation method.

2.1 Micro-aggregation

As mentioned in Sect. 1 the MAP has been tackled using different techniques. Basically, a MAT relies on a clustering technique and an aggregation technique. MATs were originally used for numerical data [4, 32], and they can be further classified as below.

2.1.1 Uni-variate versus multi-variate

The difference between the uni-variate and the multi-variate MATs depends on the number of random variables used in the micro-aggregation process. Uni-variate MATs deal with multi-variate data sets by micro-aggregating one variable at a time [7–9]. Multi-variate MATs either rank multi-variate data by projecting them onto a single axis,[4] dealing directly with the unprojected data [10, 11, 23], or using various heuristics [10, 14, 21]. More recently, researchers have advocated the use of Learning Automata [16] and Neural Networks (see [28] and the references cited there), but the details of these methods are also omitted here in the interest of brevity.

2.1.2 Fixed-size versus data-oriented

The difference between the fixed-size and the data-oriented MATs depends on the number of records in each group. Fixed-size MATs require all groups to be of size k except for a single group whose cardinality is greater than k when the total number of records, n, is not a multiple of k [10, 11, 23, 29]. Data-oriented MATs allow groups to be of size greater than k and less than $2k - 1$ depending on the structure of the data. These methods [5, 10, 17, 22, 24, 25], the details of which are omitted in the interest of brevity, yield more homogenous groups, and thus help to further minimize the IL.

2.1.3 Optimal versus heuristic

A formal algorithm to find the optimal solution for the k-partition problem was proposed by Defays and Nanopoulos [9]. But, the first reported optimal uni-variate MAT with a polynomial complexity is given in [18], which solves the MAP as a shortest path problem on a graph. Unfortunately, determining the optimal MAP for multi-variate micro-aggregation is an NP-hard problem [27]. Therefore, researchers seek heuristic MATs that provide a good solution—close to the optimal.

2.2 Maximum distance average vector

The first algorithm to accomplish micro-aggregation without projecting the multi-variate data onto a single axis was proposed in 2002 by Domingo-Ferrer and Mateo-Sanz [10], and is known as the MDAV. It micro-aggregates the multi-variate micro-data file based on the concept of the diameter distance of the data set. In 2005, an enhanced version of MDAV appeared in [14], and was implemented as a built-in technique in the μ-ARGUS software tool version 4.0 [19]. The modification is based on utilizing the centroid concept (instead of the diameter) in the micro-aggregation. In a nutshell, the process is as follows: First of all, the algorithm computes the centroid of the data. After this, a quick search for the most distant record from the centroid, say X_r, is done. Subsequently, a new search for

[2] The reader will observe that all our attention has been on minimizing the IL. This is because previous researchers in the field have also advocated such an optimization. To achieve this, as mentioned earlier, they have proposed using a subset of the variables. Of course, a more comprehensive study should also involve the DR, or a combination of the IL and the DR. This certainly leads to many unsolved problems, and we are very grateful to the anonymous Referee who suggested this.

[3] The bibliography and citations presented in this paper (for this field and for the areas covered in the next sections) were quite extensive in the earlier version of the paper. They have been abridged at the request of the Referees.

[4] The multi-variate data is projected onto a single axis by using either a particular variable, the sum-z-scores or a principle component analysis prior to micro-aggregation [24, 25].

the most distant record from the record X_r, say X_s, is accomplished. The next step consists of creating two clusters, the first one comprising of X_r and *its* $k-1$ nearest records, while the second comprises of X_s with *its* nearest $k-1$ records. At the end of this stage, the two clusters are micro-aggregated and removed from the original data set. The latter steps are iteratively repeated until there are no more records remaining in the original data set. The advantages of this new modified version of the MDAV are the increased speed of the micro-aggregation, and the reduction in the IL.

More recently, the V-MDAV scheme was proposed to obtain a data-oriented micro-aggregation solution, which provides variable-sized groups, leading to a higher within-group homogeneity while maintaining an equivalent computational cost [31].

3 Enhancing micro-aggregation with dependence

It is well-known that the result of the multi-variate MATs depends on the number and the *identity* of the variables used in the micro-aggregation process. Since multi-variate micro-aggregation using two or three variables at a time offers the best trade-off between the IL and the DR [13], the question we intend to resolve involves understanding why we have to maintain and use vast dimension-dependent resources in the clustering phase in order to compute the distance between the micro-records. We shall also study how we can minimize the computation time needed to evaluate the distance between a single micro-data record and the mean of the group it belongs to. This computation involves evaluating

$$D(X,\bar{X}) = \sqrt{\sum_{i=1}^{d}(x_i - \bar{x}_i)^2}, \tag{1}$$

where X and $\bar{X}$ are the two multi-variate data vectors (in particular, note that the second vector is the mean of the instantiations of X) whose components are $\{x_i\}$ and $\{\bar{x}_i\}$, respectively, and d represents the dimension of the space.

We consider the problem of determining the dependencies between the different variables within a micro-data file, and then combining the latter with the MAT in such a way as to reduce the overall required computational time, and/or reduce the corresponding IL.

The primary goal of any MAT is to reduce the loss in the data utility by choosing the most suitable sub-set of variables with size equal to two, three or four [13] prior to invoking the multi-variate micro-aggregate. Theoretically, to know the best sub-set of variables that has to be used in order to obtain the minimum value of the IL, we have to consider all different possibilities of combinations, namely the $\binom{S}{C} = \frac{S!}{C!(S-C)!}$ combinations, where S is the number of variables in the original micro-data file, and C is the number of chosen variables which are used in projecting and micro-aggregating the data file.

We propose that the key idea in choosing a sub-set of the variables by avoiding the combinatorial solution, should be based on the dependence model of the micro-data file. If the variables are highly correlated, then using any one of them will somehow reflect the stochastic nature of the others. If we, thus, incorporate this logic into our consideration, we believe that we can reduce the number of variables which will be used to measure either the distance between the micro-unit and the mean of the group it belongs to, or the distance between the micro-units themselves. For a truly comprehensive comparison, it can be argued that the distances used to evaluate the IL must be computed in the given d-dimensional problem space and not in the space with reduced dimensionality. However, given the context of the problem and the state-of-the-art results [10] that have motivated this work (i.e., the issue of choosing a subset of the variables to obtain almost the same IL), we believe that it is more meaningful to see how the IL is effected after the dimensionality is reduced.[5] Thus, in turn, since we reduce the dimensionality of the space to $d' < d$, the new distance that will thus be computed will be:

$$D'(X,\bar{X}) = \sqrt{\sum_{i=1}^{d'}(x_i - \bar{x}_i)^2} \quad \text{where } d' < d. \tag{2}$$

The reader should observe that our goal is quite distinct from the reported methods of projecting the multi-dimensional space onto a single axis using a particular variable, the sum z-scores scheme, or a principal component analysis. The reduction in the dimensionality is not done randomly. Rather it is to be done based on a formal criterion. Our aim is to micro-aggregate the multi-dimensional vector by maximally using the information in the "almost-independent" variables, and we plan to do this by finding the best dependence tree. We believe that we can achieve this by evaluating the dependence between the variables in the micro-data file by using either the method due to Chow and Liu [2] or the method due to Valiveti and Oommen [33, 34].

We formalize these concepts below. The joint probability distribution of the random vector $\mathbf{V} = [V_1, V_2, \ldots, V_d]^{\mathrm{T}}$ in terms of conditional probabilities is given as

[5] As mentioned earlier, we agree with the recommendation of the anonymous Referee, who suggested that a more fair comparison should involve both the IL and the DR measures.

$$P(\mathbf{V}) = P(V_1)P(V_2|V_1)P(V_3|V_1,V_2)\ldots P(V_d|V_1,V_2,\ldots,V_{d-1}), \quad (3)$$

where each V_i is a random variable.

It is obvious, from the above expression, that each variable is conditioned on an increasing number of other variables. Therefore, estimating the kth term of this equation requires maintaining the estimates of all the kth order marginals. Clearly, it is impractical to gather the estimates for the joint density function $P(\mathbf{V})$ for all the different values which $\mathbf{V}$ could assume. We, therefore, simplify the dependency model by restricting ourselves to the lower-order marginals, using the approximation which ignores the conditioning on multiple variables, and retaining only dependencies on at most a single variable at a time. This leads us to the following [33]:

$$P_{\mathrm{a}}(\mathbf{V}) = \prod_{i=1}^{d} Pr(V_i|V_{j(i)}), \quad (4)$$

where $P_{\mathrm{a}}(\mathbf{V})$ is the approximated form of $P(\mathbf{V})$, and V_i is conditioned on $V_{j(i)}$ for $0 \leq j(i) < i$.

The dependence of the variables can be represented as a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{W})$ where $\mathbf{V} = \{V_1, V_2, \ldots, V_d\}$ is a finite set of vertices, which represents the set of random variables in the micro-data file with d dimensions, $\mathbf{E}$ is a finite set of edges $\{\langle V_i, V_j\rangle\}$, where $\langle V_i, V_j\rangle$ represents an edge between the vertices V_i and V_j. Finally, $\mathbf{W} = \{w_{i,j}\}$ is a finite set of weights, where $w_{i,j}$ is the weight assigned to the edge $\langle V_i, V_j\rangle$ in the graph. The values of these weights can be calculated based on a number of measures, as will be explained presently.

In $\mathbf{G}$, the edge between any two nodes represents the fact that these variables are statistically dependent [2]. In such a case, the weight, $w_{i,j}$, can be assigned to the edge as being equal to the expected mutual information measure (EMIM) metric between them. Generally speaking, the EMIM metric between two variables, given by $I^*(V_i, Vj)$ for discrete distributions, has the form:

$$I^*(V_i, V_j) = \sum_{v_i, v_j} Pr(v_i, v_j) \log \frac{Pr(v_i, v_j)}{Pr(v_i)Pr(v_j)}, \quad (5)$$

where the summation above is done over all values of v_i and v_j which V_i and V_j can assume.

Observe that any edge, say $\langle V_i, V_j\rangle$ with the edge weight $I^*(V_i, V_j)$ represents the fact that V_i is stochastically dependent on V_j, or that V_j is stochastically dependent on V_i. Although, in the worst case, any variable pair could be dependent, the model expressed by Eq. (4) imposes a tree-like dependence. It is easy to see that this graph includes a large number of trees (actually, an $O(d^{(d-2)})$ of such spanning trees). Each of these trees represents a unique approximated form for the density function $P(\mathbf{V})$. Chow and Liu proved that searching for the best "dependence tree" is exactly equivalent to searching for the maximum spanning tree[6](MST) of the graph [2]. Further, since the probabilities that are required for computing the edge weights are not known a priori, Valiveti and Oommen showed that this could be achieved by estimating them in a maximum likelihood (ML) manner [33, 34]. They showed that the ML estimate for the best dependence tree, can be obtained by computing the MST of the graph, where the edge weights are computed using the EMIM of the estimated probabilities, as shown in Fig. 1.

It is worth mentioning that this solution is truly both elegant and efficient. A rigorous *ML* solution to obtaining the best tree would involve computing *it* from the set of all possible spanning trees, which (at the enumeration level itself) is a combinatorially explosive problem. To solve the ML problem in a formal manner, one has to first obtain the set of all the graph's spanning trees, and then determine the tree which maximizes the likelihood function evaluated in terms of the dependence described by the tree itself. Observe that the solution obtained by solving for the MST is many orders of magnitude less complex. It involves estimating *the probabilities* (and not the structure) of the Binomial (multinomial) distributions using a ML estimate, and then merely computing the MST. The fact that these two processes lead to the same estimate (as shown in Fig. 1) is far from trivial to prove, but is indeed, true.

It should be mentioned here that the weights of the edges in the graph, $\mathbf{G}$, can be computed using either the EMIM metric or the χ^2 metric proposed by Valiveti and Oommen [33]. The latter, $I_\chi(V_i, V_j)$, is an alternative measure that quantifies the dependence information between pairs of random variables, and is computed by:

$$I_\chi(V_i, V_j) = \sum_{v_i, v_j} \frac{(Pr(v_i, v_j) - P(v_i)P(v_j))^2}{P(v_i)P(v_j)}. \quad (6)$$

I_χ has the following desirable characteristics relevant to capturing dependence information:

$$\begin{cases} I_\chi(V_i, V_j) = 0 & \text{iff } P(v_i, v_j) = P(v_i)P(v_j) \\ I_\chi(V_i, V_j) > 0 & \text{otherwise.} \end{cases} \quad (7)$$

It turns out that for binary and normally distributed random variables, the I_χ metric is exactly equivalent to the I^* metric in finding the dependence tree [33, 34]. But, when the underlying dependence is not actually based on a tree

[6] Two generic greedy algorithms can be used to solve the minimum spanning tree problem, namely, the so-called Kruskal and the so-called Prim algorithms. Both of them run in time $\bigcirc(E \lg V)$ by using ordinary binary heaps [3]. Since we are attempting to compute the MST, it is obvious that we have to order the edges in a decreasing order (as in Kruskal) or to extract the maximum edges weight (as in Prim). We have used the Kruskal algorithm in our experiments.

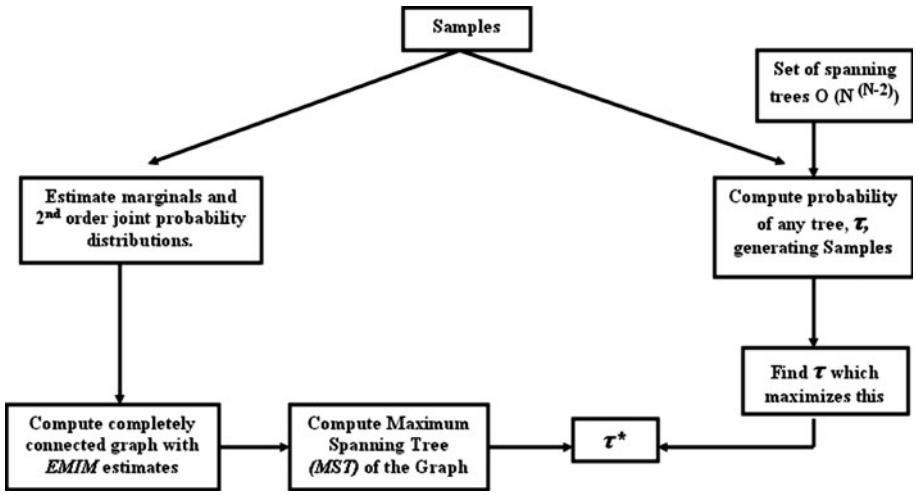


Fig. 1 Equivalent procedures for finding the maximum likelihood estimate of the tree-based dependence from the samples

structure, both of them estimate the best dependence tree corresponding to their representative measures. Valiveti and Oommen showed the interesting feature that although their estimation for the best dependence tree does not always match, the total weights are almost always identical.

By way of example, consider a micro-data file which incorporates six variables (as in Fig. 2) and thousands of records. Let us assume that we intend to micro-aggregate this file using any MAT, for example, the MDAV method. In such a case, the prior art will process all the six variables to quantify the relevant distances during the clustering stage. We could choose a sub-set of size three to be used in the micro-aggregation process. In general, we will have to go through the 20 different combinations of size three in order to attain the minimum value of the IL. However, if we are able to discover any existing inter-variable dependencies, this could render the problem simpler. Let us assume that we compute the EMIM-based edge weights for all pairs of nodes, and create the fully connected undirected graph G, as in Fig. 2. By using the strategy alluded to above, we obtain a tree as in Fig. 3a, which shows the case when the MST leads to the ML condition that the variables B, C, and D depend on the variable A, and that variables E and F depend on variable D.

Since these dependent variables are maximally correlated to the variable that they depend on, we propose to use the vertices that have the maximum number of In/Out edges in the graph to micro-aggregate the micro-file. We believe that the nodes which possess this property are the best candidates to reflect the characteristics of the entire multi-variate data set because they connect to the maximum number of nodes that statistically depend on it, as argued in Conjecture 1.

Conjecture 1. Micro-aggregating the micro-data file can be best achieved if the nodes which possess the maximum number of In/Out edges in the tree obtained as the MST of the underlying undirected connected graph **G**, are used as the input to solve the MAT.

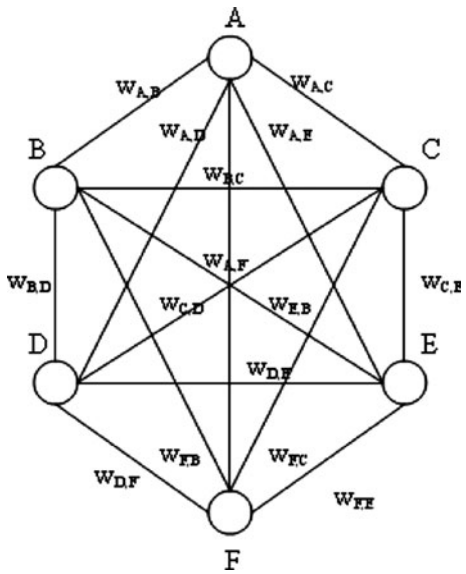


Fig. 2 The fully connected undirected graph represents the dependence between six random variables

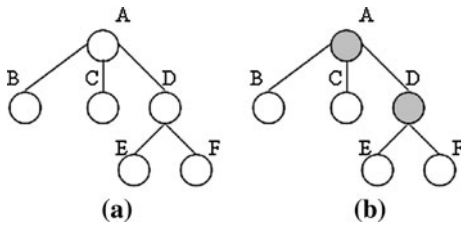


Fig. 3 An example of a dependence tree used to micro-aggregate the data file containing six variables

Rationale for conjecture The existence of an edge between two nodes in the connected undirected graph

signifies that these two nodes are statistically correlated to each other, and that a variation of one of these variable is reflected by a corresponding change in the other. Thus, the variables which are connected to each other via edges in the skeletal tree represent nodes which are connected to each other based on the best tree-based dependence, and in turn, reflect the maximal shared characteristics within the variables of the micro-data file. Thus, any node which has a larger number of In/Out edges is one which connects to a larger number of nodes, and is thus capable of individually representing more "other" variables. This implies that the best candidates to be used to represent the other variables in the micro-aggregation are those which have the maximum number of In/Out edges.

In order to invoke this property, we first rank the nodes of the graph based on the number of In/Out edges in a descending order and choose the first d' variables, where d' is usually determined by the data protector, and is usually equal to 3 or 4. Thus, for example, based on the above discussion, for the data represented by the variables of Fig. 3, the micro-aggregation process will be invoked by using two variables instead of using the entire set of six variables in the micro-data file. Figure 3b shows that the selected sub-set of the variables is $\{A, D\}$, since both of them connect to three variables while the other variables in the micro-data file connect to only a single variable. The process outlined above has been formalized in Algorithm 1 which presents an automated way to select a sub-set of the variables to be used in the multi-variate micro-aggregation process.

Table 1 The characteristics of various data sets

Name of the data set	Type	Dimensionality	Cardinality
Tarragona	Real	13	834
Census	Real	13	1,080
*Sim*_1	Simulated	8	5,000
*Sim*_2	Simulated	16	10,000
*Sim*_3	Simulated	22	20,000

4 Experimental result

4.1 Data sets

In order to verify the validity of our methodology in projecting the multi-variate data set into a subset of random variables to be used in the micro-aggregation process, two benchmark real-life data sets and three simulated data sets were used in the testing phase. Table 1 summarizes the characteristics of each data set by defining its type, dimensionality and cardinality.

The Tarragona and Census benchmarks are reference data sets used in previous studies for their special statistical properties [10, 12]. On the other hand, the simulated data were generated or tested for various dimensions of random vectors, as follows: first of all, the number of random variables was determined. Thereafter, the "true" structure of the defined dependence tree which imposed the dependence relationships between the variables was selected subjectively, as shown in Fig. 4. Then the second-order marginal distributions were randomly generated. The procedure by which these were generated was as follows: if we

Algorithm 1 *EMAD*

Input: $\mathcal{U}$: the micro-data file, and C: the number of variables that will be used in the micro-aggregation process.
Output: d': the sub-set of the variables that will be used in the multi-variate MAT.
Method:

1: Estimate the first and second order marginals of the random variables from the various micro-records.
2: Create a fully-connected undirected graph, where the
Nodes: Represent the random variables in the micro-data file.
Edges: Represent the possible statistical dependence between these variables.
Weights of the edges are computed either by using:
$EMIM \Rightarrow I^*(V_i, V_j) = \sum_{v_i,v_j} Pr(v_i, v_j) \log \frac{Pr(v_i,v_j)}{Pr(v_i)Pr(v_j)}$, OR
$\chi^2 \Rightarrow I_\chi(V_i, V_j) = \sum_{v_i,v_j} \frac{(Pr(v_i,v_j)-P(v_i)P(v_j))^2}{P(v_i)P(v_j)}$.
3: Invoke Kruskal's algorithm to compute the Maximum Spanning Tree of the graph.
4: Rank the nodes of the graph based on the number of In/Out edges in a decreasing order, and reckon the first d' variables to be the sub-set to be used in the MAT.
5: **Return** the sub-set of variables which will be used in the micro-aggregation process before invoking the MAT
6: **End Algorithm** *EMAD*

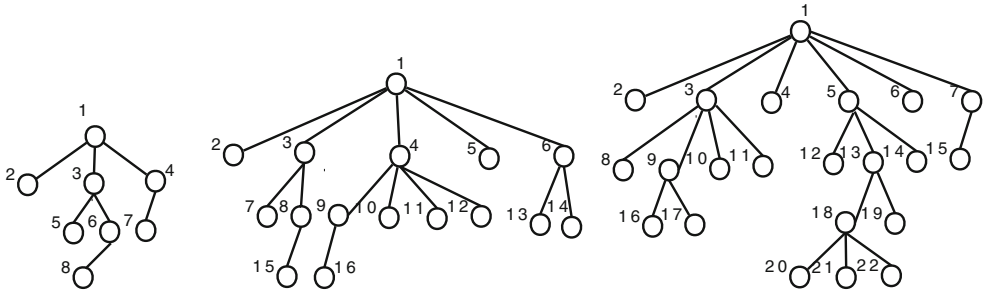


Fig. 4 The true structures for the simulated data sets

define the entire space of each variable to be between 1 and 1,000, this space is sub-divided into a number of subspaces with equal width, say 100. That means that we limit ourselves to be dealing with ten events where each event represents a sub-interval of width equal to 100 from the entire domain as follows: $\{I_1 = [1, 100], I_2 = [101, 200], \ldots, I_{10} = [901, 1{,}000]\}$, thus, effectively simulating a multinomial distribution. In the latter, each outcome is a random number belonging to exactly one of the ten sub-intervals, I_j, with probability, P_j, where $j = 1, 2, \ldots, 10$. If n_j represents the number of occurrences of values belonging to I_j and n represents the number of independent records, we have

$$\sum_{i=1}^{10} n_i = n, \sum_{i=1}^{10} P_i = 1, \tag{8}$$

where the probability mass function of the multinomial distribution is

$$f(n_1, n_2, \ldots, n_{10}) = \frac{n!}{n_1! n_2! \ldots n_{10}!} \prod_{i=1}^{10} P_i^{n_i}. \tag{9}$$

Observe that prior to assigning the second-order marginal distributions for the rest of the tree, we had to also randomly generate ten different probabilities for the most independent variables (the root variable) when its values belonged to each of the above defined sub-intervals.

To randomly populate the file, we can now randomly assign values to the conditional probability from the joint and marginal distributions as follows: if, as per the assumed tree-based dependence, variable V_m, depends on variable V_n, this means we have to define a set of probabilities, $\{P_{nm}\}$, when the value of V_n, say v_{in}, belongs to any defined sub-interval I_j given that the value of variable V_m, say v_{im} belongs to any sub-interval I_l. Thus,

$$P_{mn} = Pr(v_{in} \in I_j | v_{im} \in I_l), \tag{10}$$

where i represents the index of the record in the micro-data file and assumes values in $\{1, 2, \ldots, n\}$. The indices j and l represent the indices of the sub-interval where the random variable falls, and which are the result of dividing the entire domain into ten sub-intervals. Finally, the indices n and m represent the specific dimensions in the micro-data file, and are in the range $\{1, \ldots, d\}, n \neq m$.

The above procedure was implemented for all pairwise combinations of random variables associated with the micro-data file.

4.2 Results

The experiments conducted fell into four categories, where in each case[7] the value of k was set to 3: in the first set of experiments the intention was primarily focused on testing whether the best dependence tree can be learned (or rather, inferred) from the continuous micro-data file, and if it sufficiently reflected the dependence model. In the second set of experiments, the goal was primarily to validate our strategy for determining the subset of variables (from the entire set of variables) to micro-aggregate the micro-data file, and to study its effect on the value of the IL. The third set of experiments was designed to determine the most suitable metric to calculate the edge weights of the fully connected graph so as to minimize the required computation time and maximize the accuracy of estimating the dependence model and its effect on the value of the IL. Finally, since we are working with continuous vectors, the last set of experiments focused on understanding the effect of assuming normality (i.e., the relevance of the Central Limit Theorem [15]) on the data set in calculating the edges weights.

[7] Throughout this section, we have, in the interest of brevity, only reported the results for the case when $k = 3$. This is because researchers who have worked with MATs have advocated setting $k = 3$ or 4 independent of the dimension of the multivariate vector. Observe that once the value of k has been set, the difference between the IL in the original space and the reduced space is of primary importance. Our experience is that the respective difference between the IL (in the original and reduced subspaces) for the cases when $k = 3$ and $k = 4$ is minimal.

4.2.1 Experiment sets 1

The first set of experiments was done on two types of data sets: simulated data sets with a known structure of the best dependence tree which is to be inferred by the learning algorithm, and the real data sets possessing an unknown dependence model between the variables. It is worth mentioning that we could not approximate the dependence information of the multi-variate data set in its current form due to the inaccurate estimation for the joint and marginal probability distributions for continuous variables. This is a consequence of having a large domain space with only few records (sometimes only 1 or 2) for each region of the corresponding random variable. Consequently, most of the *estimated* marginal and joint probability values were close to zero. Clearly, in these cases, the *estimated* probabilities will not reflect the actual dependence relationship between any corresponding variables.

In order to overcome this challenging problem that prevents us from utilizing the dependence information, we were forced to reduce the domain space by categorizing the micro-data file as follows: we first scanned the micro-data file to specify the domain space of each variable in the file, and then divided it into a number of sub-intervals sharing the same width. After that, we achieved a categorization phase by replacing the values belonging to a certain sub-interval in each variable by the corresponding category/code. For example, in the case of the simulated data sets, all the variables shared the same domain space between 1 and 1,000, which was divided into ten subintervals, as explained earlier. Consequently, all values belong to the [1,100] interval were replaced by 1, all values belong to the [101,200] interval were replaced by 2 and so on. The above procedure was repeated for all the variables so as to generate the categorical micro-data file.

From the above discussion, it is clearly shown that "width" parameter plays a predominant role in controlling the degree of smoothing and estimating the best dependence tree. Our experiments indicated that assigning a suitable value to the width parameter guaranteed the convergence of the MST to the true underlying (unknown) structure of the best dependence tree. The most important point that one has to be aware of in a practical scenario is that a larger value for the width parameter implies a lower variance and a higher bias, because we are essentially assuming a constant value within the sub-interval. Generally speaking, the value of the width parameter should be large enough to generate a sufficient number of sub-intervals from the defined domain space to guarantee a satisfactory level of smoothing. The actual value used is specified in the respective experimental results.

Consider the tree structure given by *Sim*_1, *Sim*_2, and *Sim*_3 as given in Fig. 4. Approximating the dependence information of the simulated data sets based on the structure of the MST obtained using the EMIM metric succeeded in locating the real structure when the width parameter was set to the values 50,100, and 150 for *Sim*_1, 70, 100, and 120 for *Sim*_2, and 90, 100 and 110 for *Sim*_3. Figure 5 shows the edge weights and the value of I_x for each simulated data set when the value of width was equal to 100. Figures 6, 7 and 8 show different snapshots of the convergence to the final structure of the dependence model for various sample sizes for *Sim*_1, *Sim*_2 and *Sim*_3, respectively, when the value of the width parameter was set to 100.

Approximating the dependence information for the real data sets was a little more "tricky", because of the unknown structure for the best dependence tree. Changing the value of the width parameter has an effect on the structure of the best dependence tree to which the algorithm converged. Figures 9 and 10 clearly show different structures for the best dependence tree by changing the value of the width for the Tarragona and Census data sets, respectively.

The final set of experiments involves the so-called *Sibling-related* Model. The aim here was to see if the algorithms possessed the ability to infer the structure of the dependence model between the random variables if additional information about the dependency between the siblings in the tree is available. The results that we have obtained are quite amazing.

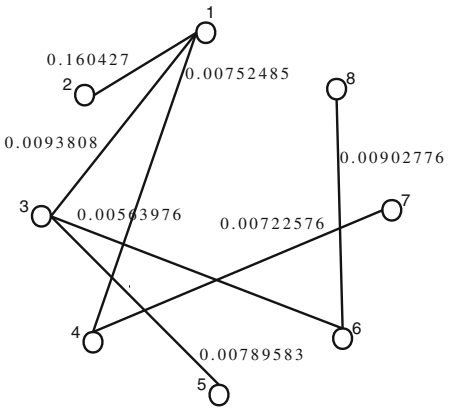
To be more specific, we consider the possibility that after the structure of the underlying tree is determined, the *probability* values between the siblings in the structural tree are related. For example, thus, if a particular node had index i and its children were nodes j and k, the probabilities that could be *independently* set were:

$$Pr[x_j = 0|x_i = 0]$$

$$Pr[x_j = 0|x_i = 1].$$

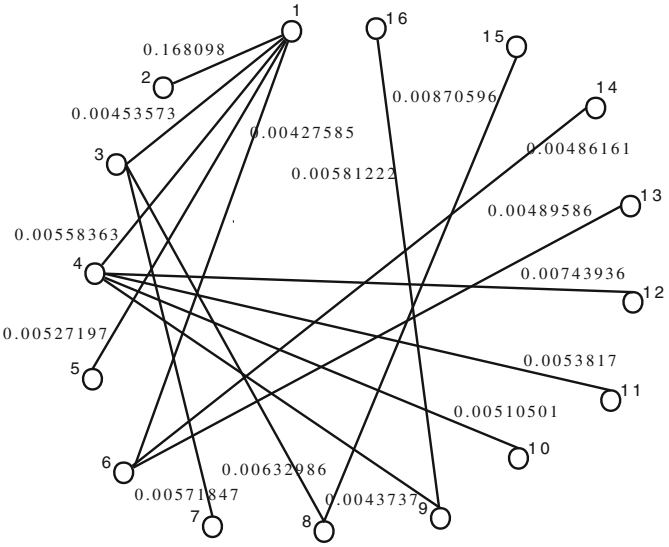
Since the probabilities of the siblings were thus determined, the values of $Pr[x_k = 0|x_i = 0]$ and $Pr[x_k = 0|x_i = 1]$ were then set to be $1 - Pr[x_j = 0|x_i = 0]$ and $1 - Pr[x_j = 0|x_i = 1]$, respectively. Further, observe that a result of these assignments, the probabilities, $Pr[x_j = 1|x_i = 0], Pr[x_j = 1|x_i = 1], Pr[x_k = 1|x_i = 0]$ and $Pr[x_k = 1|x_i = 1]$ were automatically assigned, since the sum of these quantities and the values of their counterparts, is unity.

The question we were interested in investigating was to see if our strategy for learning the dependence tree using the MST on the constructed fully connected graph (where the edges weights are calculated using the EMIM or the χ^2 metric) was able to converge to the true (unknown) dependence tree even if this sibling relationship was not known. The answer was always in the affirmative.



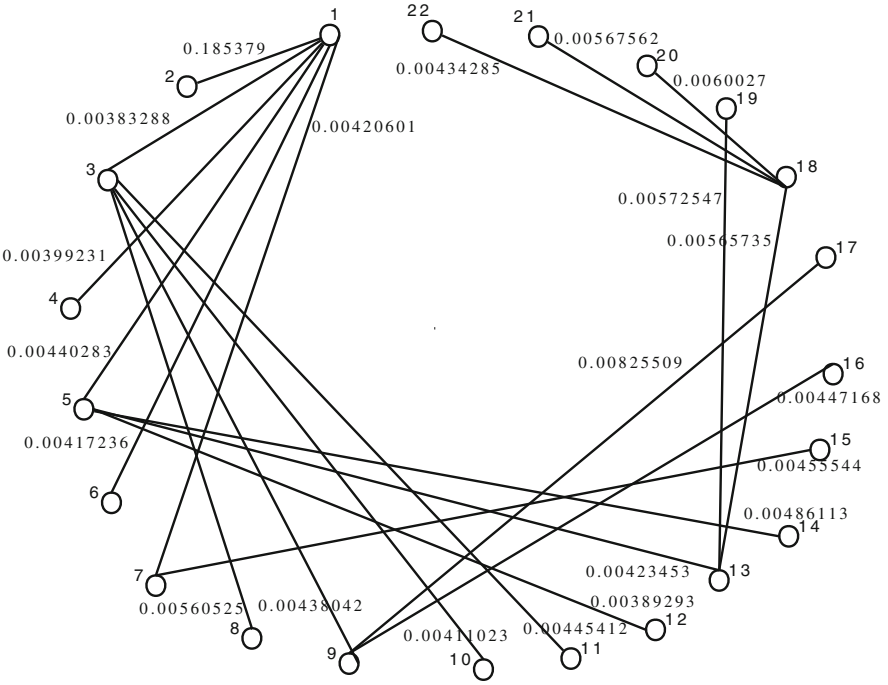
The Total Weight = 0.207122

(a) Sim_1



The Total Weight = 0.246389

(b) Sim_2

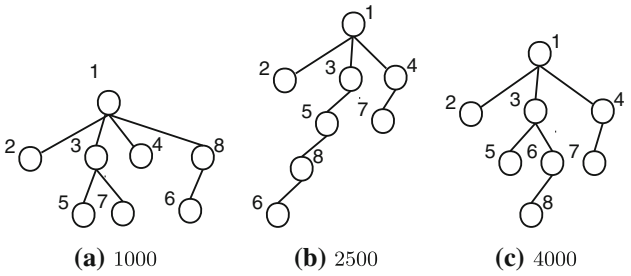


The Total Weight = 0.282211

(c) Sim_3

Fig. 5 The best dependence tree for the simulated data sets obtained by using the EMIM metric

By way of example, consider two binomial data sets with six variables. Both of them share the same dependency model between the variables, as shown in Fig. 11. The only difference between these two data sets is that the values of the probabilities used to generate the random variables in the true tree structure—which in one case *was* sibling-related, and in the other *was not* sibling-related. Tables 2 and 3 show the values of the random probabilities which were used in generating each variable in the data set. Observe that in the first data set these values are related,



(a) 1000 **(b)** 2500 **(c)** 4000

Fig. 6 The "inferred" dependence tree for the *Sim*_1 binary data set as the number of samples increases. The width parameter was set to 100

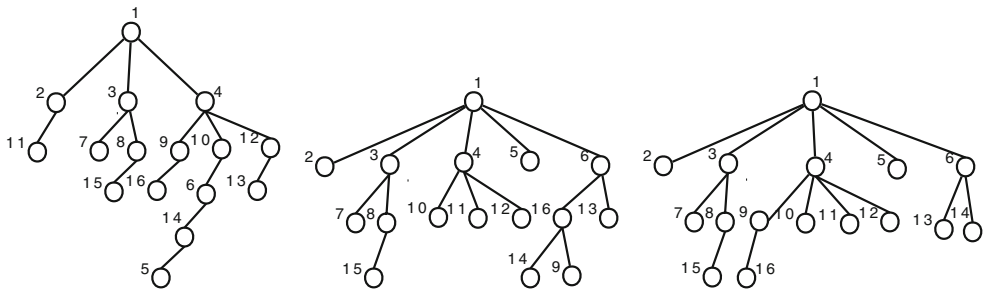
while they are independent in the second data set. Figure 12 show different snapshots of the convergence to the dependence model as the number of samples is increased.

The actual trees learnt for the data sets, as the number of samples processed increased, are given in Figs. 12 and 13, respectively (reported at snapshots 50, 150 and 5,000). The decrease in the EMIM and χ^2 metrics with time are plotted in Fig. 14. Observe that the final inferred tree in both cases is exactly the unknown tree—which, again, was correctly inferred, and that the values of both the metrics ultimately converged to the lowest possible values. Thus we conclude that the relationship between the probabilities of generation of the sibling random variables, was not able to "confuse" the algorithm in learning the unknown structure.

It should be mentioned, though, that in the cases in which the sibling probabilities were related, the learning was faster—which we believe is quite remarkable.

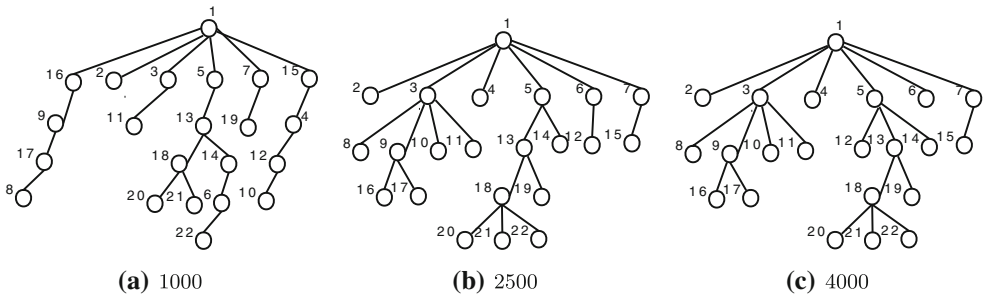
4.2.2 *Experiment sets 2*

The second set of experiments verified our conjecture that it was expedient to use the sub-set of the variables obtained (from the best dependence tree) by projecting the micro-



(a) 1000 **(b)** 2500 **(c)** 4000

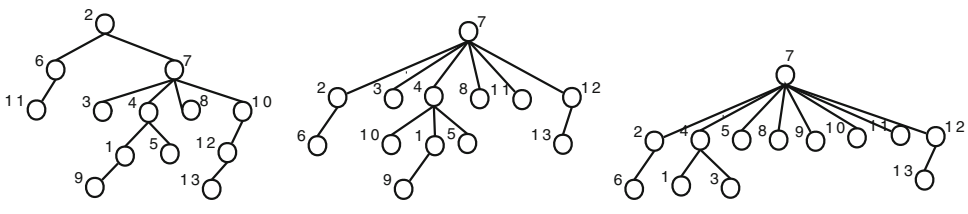
Fig. 7 The "inferred" dependence tree for the *Sim*_2 binary data set as the number of samples increases. The width parameter was set to 100



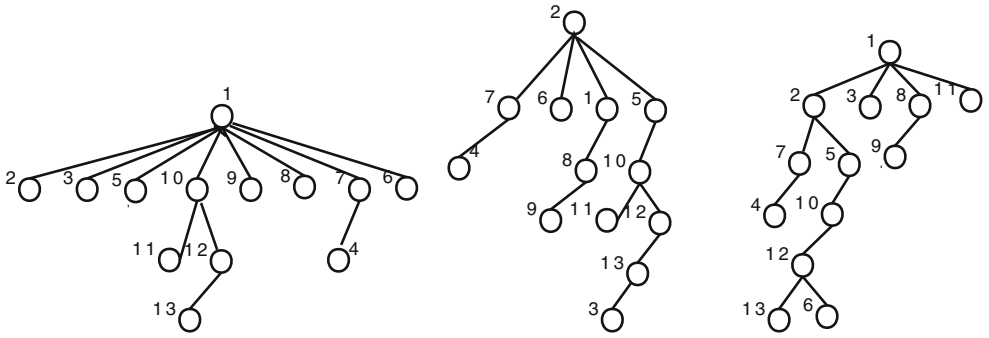
(a) 1000 **(b)** 2500 **(c)** 4000

Fig. 8 The "inferred" dependence tree for the *Sim*_3 binary data set as the number of samples increases. The width parameter was set to 100

Fig. 9 The best dependence tree for the Tarragona data set obtained by using the EMIM metric with various values of the width parameter



(a) Width $= 50,000$ **(b)** Width $= 100,000$ **(c)** Width $= 150,000$



(a) Width $= 1,000$ **(b)** Width $= 5,000$ **(c)** Width $= 10,000$

Fig. 10 The best dependence tree for the Census data set obtained by using the EMIM metric with various values of the width parameter

Fig. 11 The best dependence tree for a binomial data sets with 5,000 records and six variables

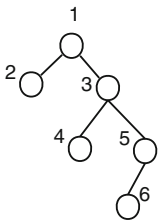
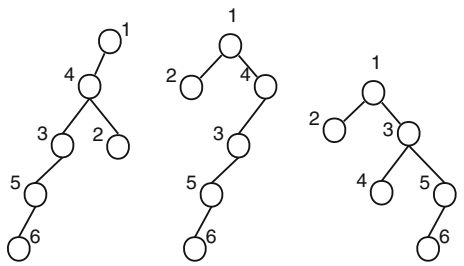


Table 2 The probability values used in generating the corresponding random variables when the corresponding probabilities for the sibling nodes in the structural dependence tree are related

Probability	Value
Prob($x1 = 0$)	0.40
Prob($x1 = 1$)	0.60
Prob($x2 = 0 \| x1 = 0$)	0.30
Prob($x2 = 0 \| x1 = 1$)	0.10
Prob($x3 = 0 \| x1 = 0$)	0.70
Prob($x3 = 0 \| x1 = 1$)	0.90
Prob($x4 = 0 \| x1 = 0$)	0.20
Prob($x4 = 0 \| x1 = 1$)	0.60
Prob($x5 = 0 \| x3 = 0$)	0.80
Prob($x5 = 0 \| x3 = 1$)	0.40
Prob($x6 = 0 \| x3 = 0$)	0.15
Prob($x6 = 0 \| x3 = 1$)	0.76

Table 3 The probability values used in generating the corresponding random variables when the corresponding probabilities for the sibling nodes in the structural dependence tree are unrelated

Probability	Value
Prob($x1 = 0$)	0.40
Prob($x1 = 1$)	0.60
Prob($x2 = 0 \| x1 = 0$)	0.30
Prob($x2 = 0 \| x1 = 1$)	0.10
Prob($x3 = 0 \| x1 = 0$)	0.60
Prob($x33 = 0 \| x1 = 1$)	0.70
Prob($x4 = 0 \| x1 = 0$)	0.20
Prob($x4 = 0 \| x1 = 1$)	0.60
Prob($x5 = 0 \| x3 = 0$)	0.40
Prob($x5 = 0 \| x3 = 1$)	0.50
Prob($x6 = 0 \| x3 = 0$)	0.15
Prob($x6 = 0 \| x3 = 1$)	0.76



(a) $N = 50$ **(b)** $N = 70$ **(c)** $N = 80$

Fig. 12 The "inferred" dependence tree for the binary data set as the number of samples increases. In this case, the probabilities between the sibling random variables are *related*

data file into three, four or five variables before invoking the multi-variate micro-aggregation process.

Since an MAT seeks to reduce the loss in the data utility, it must be pointed out here that the value of the IL depends on the sub-set of variables used to micro-aggregate the multi-variate data file. As mentioned earlier, to infer the best sub-set of variables to be used in the micro-aggregation, we have to go through all the different projection possibilities. The results (Table 4) show that the estimation of the percentage value of the IL for various data sets obtained by projecting the entire data set into specified number of variables prior to invoking the MDAV method, for which the value of k was again set to 3. The value of the IL was bounded between the minimum value (in the fourth column) that was obtained by using the variable indices addressed in the third column, and the maximum value (in the sixth column) that was obtained by using the indices addressed in the fifth column. The last column in Table 4

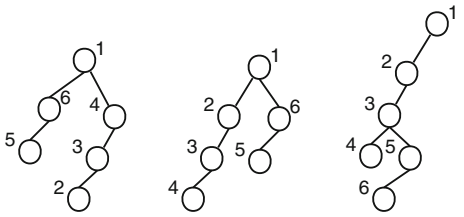


Fig. 13 The "inferred" dependence tree for the binary data set as the number of samples increases. In this case, the probabilities between the sibling random variables are *unrelated*

represents the average value of the IL over all the different combinations of projected variables in the micro-data file.

The most interesting observation was that the minimum value of the IL obtained by using three, four or five projected variables in the Tarragona and Census data sets were exactly the same. This implies using the same "most independent variables", which in turn, preserve the same high value for the variance. Therefore, in the case of real-life data sets, we recommend projecting the entire micro-data file using three variables, since using a larger number of variables to project the micro-data file requires more

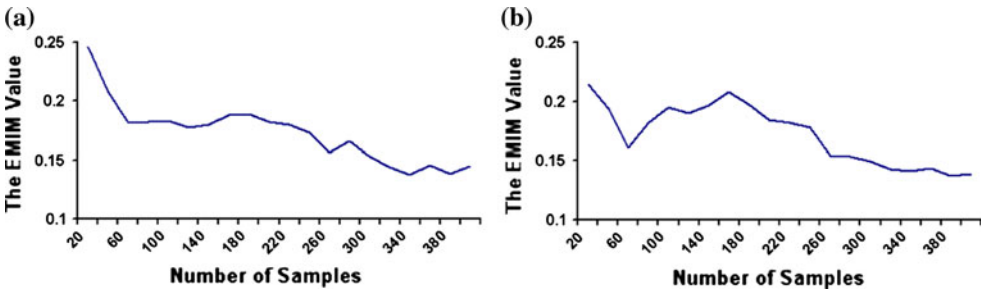


Fig. 14 The convergence of the corresponding metric for the *Set-Up* four data sets by using **a** the EMIM metric to calculate the edges weights. **a** The probabilities between the siblings are *related*, and **b** these probabilities between the siblings are *unrelated*

Table 4 The value of the IL obtained by using the MDAV multi-variate MAT after projecting various data sets into the specific number of variables

Dataset	No. of projected variables	No. of possible combinations	Indices of the variables used to obtain the min. value of IL	The min. value of IL	Indices of the variables used to obtain the max. value of IL	The max. value of IL	The average value of IL
Tarragona	1	13	10	37.6374	8	48.1006	43.1017
	2	78	11,13	24.7415	5,8	45.6925	31.6609
	3	286	2,3,10	20.7141	5,6,11	34.1569	25.1587
	4	715	2,3,10,11	20.7141	5,6,11,12	34.1569	25.4997
	5	1,287	2,3,10,11,12	20.7141	5,6,11,12,13	34.1569	25.6141
Census	1	13	10	38.2133	1	62.9093	45.79787
	2	78	4,13	22.5795	1,8	55.634	31.618
	3	286	7,8,10	15.6043	1,8,9	45.815	21.2046
	4	715	7,8,10,11	15.6043	1,8,9,10	45.815	22.0308
	5	1,287	7,8,10,11,12	15.6043	1,8,9,10,11	45.815	22.8299
*Sim*_1	1	8	2	56.6216	3	59.8342	58.6944
	2	28	1,8	46.8576	3,5	51.3179	49.4823
	3	56	2,4,7	37.6486	4,6,7	42.2961	39.7095
	4	70	2,4,7,8	37.6486	4,6,7,8	42.2961	39.5901
	5	56	1,3,5,6,7	37.6522	3,4,6,7,8	42.1577	39.7102
*Sim*_2	1	16	2	61.6118	11	63.0976	62.7308
	2	120	1,16	56.6698	9,13	59.0037	58.2026
	3	560	1,8,11	51.6551	6,8,9	54.3367	53.3249
	4	1,820	1,8,11,12	51.6551	6,8,9,10	54.3367	53.18
*Sim*_3	1	22	2	62.5849	7	64.0993	63.7251
	2	231	2,22	59.0211	8,11	60.9375	60.4842
	3	1,540	2,7,13	55.5274	4,6,14	57.7998	56.9827

time without leading to significant reduction in the IL value.

Practically, due to the exponential number of combinations, we could not cover the entire solution space so as to reach to the best sub-set of the variables to be used in the micro-aggregation[8]. As opposed to this, by involving only the vertices that have the maximum number of *I*/*O* edges in the connected undirected graph to micro-aggregate the micro-data file, we were able to obtain an acceptable value of the IL close to its lower bound, and which is always (in all the cases) superior to the average value. Thus, such an automated strategy for projecting the multi-variate data sets will reduce the solution space to be searched which, in turn, reduces the computation time required to test the candidate variables, and to choose the best sub-set from them.

Tables 5 and 6 show the percentage value of the IL obtained by using our strategy in projecting the micro-data file into sub-sets of sizes 3 and 4, respectively, prior to invoking the MDAV method (for which the value of k was again set to 3). When the Census data set was projected onto a number of variables prior to the micro-aggregation, the minimum values of the IL were equal to 17.47% when the width value was equal to 1,000 and the number of variables was set to 3 or 4, to 16.23% when the width value was equal to 5,000 and the number of variables was equal to 3 or 4. The value of the minimum IL was equal to 18.29 and to 17.70% when the width value was equal to 10,000 and the projection was onto three and four variables, respectively. It is worth mentioning that the values obtained were quite close to the lower bound of the IL, i.e., 15.60%, as shown in Table 4, besides being superior to the average values over all the different combinations (i.e., 21.20 and 22.03% for 3 and 4 variables, respectively). Similar results were obtained for the Tarragona data set when the minimum value of the IL using 3 or 4 variables was equal to 24.13% by setting the width value to 50,000 or 100,000. But, it was equal to 25.05% when the width was 150,000. Again, these values were closer to the lower bound of the IL which was 20.71%, and were superior to the average value which was close 25.5%. In Tarragona data set, the minimum values of the IL, when the width value was set to 50,000, 100,000 and 150,000, were equal to 24.13, 24.13 and 25.04%, respectively. The values obtained were quite close to the lower bound of the IL, i.e., 20.71%, as shown in Table 4, besides being superior to the average values over all the different combinations (i.e., 25.16%). Finally, we would like to state that the simulated data set yielded similar results to those of the real data sets where the minimum values of the IL were equal to 38.11% for *Sim*1, 51.95% for *Sim*2 and 55.82% for *Sim*3. These values were quite close to the lower bound of the IL which were equal to 37.64% for *Sim*1, 51.65% for *Sim*2 and 55.52% for *Sim*3, respectively.

Table 5 The value of the IL obtained by using the MDAV multivariate MAT after projecting various data sets using three variables by using the EMIM metric to calculate the edge weights in the connected undirected graph

Data set	Width value	No. of possibilities	Variable indices	IL
Tarragona	50,000	5	7,4,1	24.1333
			7,4,10	24.1881
			7,4,2	25.0465
			7,4,12	25.6574
			7,4,6	25.6826
	100,000	3	7,4,1	24.1333
			7,4,2	25.0465
			7,4,12	25.6574
	150,000	2	7,4,2	25.0465
			7,4,12	25.6574
Census	1,000	2	1,10,7	17.4700
			1,10,12	25.3632
	5,000	6	2,10,8	16.2332
			2,10,5	17.3421
			2,10,1	17.7012
			2,10,13	21.0694
			2,10,7	21.1128
			2,10,12	21.5828
	10,000	1	1,2,12	18.2996
*Sim*_1	100	2	1,3,4	51.9684
			1,3,6	52.1126
*Sim*_2	100	2	1,3,4	51.9684
			1,3,6	52.1126
*Sim*_3	100	2	1,3,5	56.1318
			1,3,18	55.8246

4.2.3 Experiment sets 3

The third set of experiments compares the EMIM and χ^2 metrics in calculating the edge weights in the connected undirected graph. Generally speaking, the χ^2 is faster in leading to a convergence to the best dependence tree than the EMIM metric since it required a smaller number of observations or records to converge. It is worth mentioning, though, that both metrics converged to the same true structure of the dependence model for the simulated data sets by setting the value of the width parameter to 100. The scenario is completely different for the real data sets, as seen in Figs. 15 and 16 which display different structures

[8] On our processor, it took up to a few hours or even days depending on the dimensionality and cardinality of the data set, to exhaustively search the entire space.

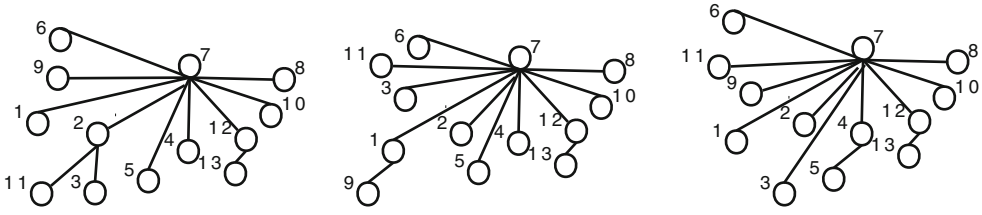
Table 6 The value of the IL obtained by using the MDAV multivariate MAT after projecting various data sets using four variables by using the EMIM metric to calculate the edge weights in the connected undirected graph

Data set	Width value	No. of possibilities	Variable indices	IL
Tarragona	50,000	10	7,4,1,10	24.1333
			7,4,1,12	24.1333
			7,4,10,12	24.1881
			7,4,1,6	24.2114
			7,4,1,2	24.9648
			7,4,2,10	25.0465
			7,4,2,12	25.0465
			7,4,6,10	25.6826
			7,4,6,12	25.6826
			7,4,2,6	26.0992
	100,000	3	7,4,1,12	24.1333
			7,4,1,2	24.9648
			7,4,2,12	25.0465
	150,000	1	7,4,2,12	25.0465
Census	1,000	1	1,10,7,12	17.4700
	5,000	15	2,10,8,12	16.2322
			2,10,8,13	16.2322
			2,10,5,8	17.0012
			2,10,7,8	17.0012
			2,10,5,12	17.3421
			2,10,5,13	17.3421
			2,10,1,12	17.7012
			2,10,1,13	17.7012
			2,10,1,5	19.4846
			2,10,7,12	21.1128
			2,10,7,13	21.1128
			2,10,12,13	21.5828
			2,10,1,8	21.9116
			2,10,7,5	23.0105
			2,10,1,7	26.4757
	10,000	4	1,2,12,10	17.7012
			1,2,12,5	19.4846
			1,2,12,8	21.9116
			1,2,12,7	26.4757
*Sim*_1	100	1	1,3,4,6	38.1105
*Sim*_2	100	1	1,3,4,6	51.9684
*Sim*_3	100	1	1,3,5,18	56.1318

for the best dependence tree for the Tarragona and Census data sets, respectively, using various values for the width parameter, and when k was 3. Table 7 shows the value of the IL obtained by invoking the MDAV method after projecting various data sets into three variables by using the χ^2 metric to calculate the edges weights in the connected undirected graph. In the simulated sets, the χ^2 metric led to the same value of the IL which was obtained by using the EMIM metric because they converged to the same dependence tree, implying that they used the same set of variables to micro-aggregate the micro-data file. As opposed to this, in the real data sets, the χ^2 converged to a different "best" dependence tree compared to the one obtained by using the EMIM metric, thus leading to a different value of the IL. In general, the value of the IL obtained by using the χ^2 metric was lower than the corresponding value obtained by using the EMIM metric for the Census data sets, but it was higher than the value obtained by using the EMIM metric in Tarragona data set. Table 7 shows that the values of the IL for the Tarragona data set, when k was 3, and the width value was set to 50,000, 100,000 and 150,000, were equal to 25.1, 24.8 and 25.7%, respectively, and for the Census data set the minimum values of the IL were equal to 17.47% when the width value was set to 1,000, 16.23% when the width value was set to 5,000, and to 18.16% when the width value was set to 10,000. In general, the χ^2-based solution space was superior to the EMIM-based solution.

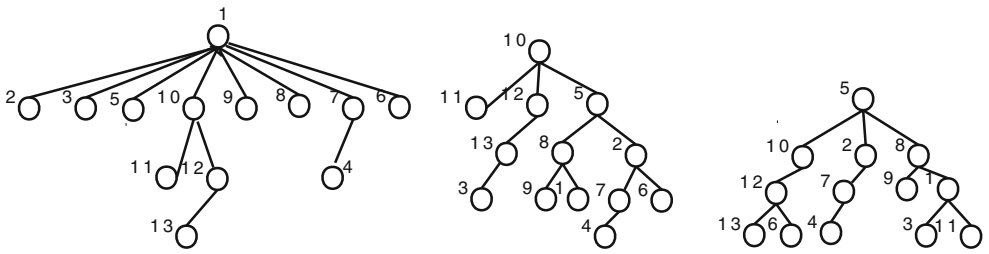
4.2.4 Experiment sets 4

The distribution of the average of a set of random variables tends to be Normal, even when the distribution from which the individual random variable is computed is decidedly non-Normal. This is a consequence of the Central Limit Theorem, which is the foundation for many statistical procedures, because the distribution of the phenomenon under study does not necessarily have to be Normal. Therefore, the last set of experiments assumes the Normality of the micro-data file to quickly compute the first and second-order marginals, and to thus lead to the MST



(a) Width $= 50,000$ **(b)** Width $= 100,000$ **(c)** Width $= 150,000$

Fig. 15 The best dependence tree for the Tarragona data set obtained by using the χ^2 metric with various values of the width parameter



(a) Width $= 1,000$ **(b)** Width $= 5,000$ **(c)** Width $= 10,000$

Fig. 16 The best dependence tree for the Census data set obtained by using the χ^2 metric with various values of the width parameter

Table 7 The value of the IL obtained by using the MDAV multivariate MAT after projecting various data sets using three variables by using the χ^2 metric to calculate the edge weights in the connected undirected graph

Data set	Width value	No. of possibilities	Variable indices	IL
Tarragona	50,000	1	7,2,12	25.0923
	100,000	1	7,1,12	24.8137
	150,000	1	7,4,12	25.6574
Census	1,000	2	1,10,7	17.4700
			1,10,12	25.3632
	5,000	4	10,2,8	16.2322
			5,8,2	17.0012
			10,5,2	17.3421
			10,5,8	22.5430
	10,000	4	1,5,8	18.1627
			1,8,12	18.8725
			1,5,12	20.2778
			5,8,12	21.9324
*Sim*_1	100	2	1,3,4	51.9684
			1,3,6	52.1126
*Sim*_2	100	2	1,3,4	51.9684
			1,3,6	52.1126
*Sim*_3	100	2	1,3,5	56.1318
			1,3,18	55.8246

for computing the best dependence tree. Subsequently, we applied our strategy to choose the subset of random variables to project the file before invoking the MDAV method (Table 8).

The beauty of estimating the dependence model assuming normality is that it does not depend on any parametric value. Therefore, it leads to a unique MST if the edges weight are unique. Figure 17 shows the best dependence tree for the simulated and real data sets for $k = 3$. It is worth mentioning that using the correlation between two random variables in calculating the edges

Table 8 The value of the IL obtained by using the MDAV multivariate MAT after projecting various data sets into three variables assuming normality

Data set	No. of possibilities	Variable indices	IL
Tarragona	5	2,4,10	23.1068
		2,4,1	24.9648
		2,4,7	25.0465
		2,4,12	25.9818
		2,4,6	26.0992
Census	5	4,10,5	16.3416
		4,10,9	16.8352
		4,10,6	19.6712
		4,10,7	20.2959
		4,10,12	20.9725
*Sim*_1	6	7,2,3	37.8147
		7,1,3	37.8676
		7,1,6	38.1610
		7,2,6	38.1704
		7,3,6	41.4843
		7,1,2	42.0634
*Sim*_2	3	4,1,3	51.9684
		4,1,12	52.0159
		4,3,12	53.9267
*Sim*_3	10	20,1,5	55.6419
		20,2,13	55.6631
		20,2,5	55.8020
		20,2,3	55.8722
		20,1,3	55.9010
		20,1,13	55.9474
		20,5,13	57.1943
		20,3,5	57.3903
		20,3,13	57.4192
		20,1,2	57.4770

weights of the graph does not lead to convergence to the "true" underlying dependence model in the case of the simulated data sets. However, generally the overall process yielded a value of IL close to the minimum value of the IL

Table 9 Characteristics of the EMIM, χ^2 and correlation metrics in calculating the edges weights of the connected undirected graph

	EMIM	χ^2	Correlation
Width parameter	Sensitive	Sensitive	Not sensitive
No. of combinations in search space	Medium	Small	Large
Convergence to the best dependence tree structure	Converge	Converge	Does not always converge
Convergence speed	Slower than χ^2 metric	Slower than assuming Normality	Faster than both metrics

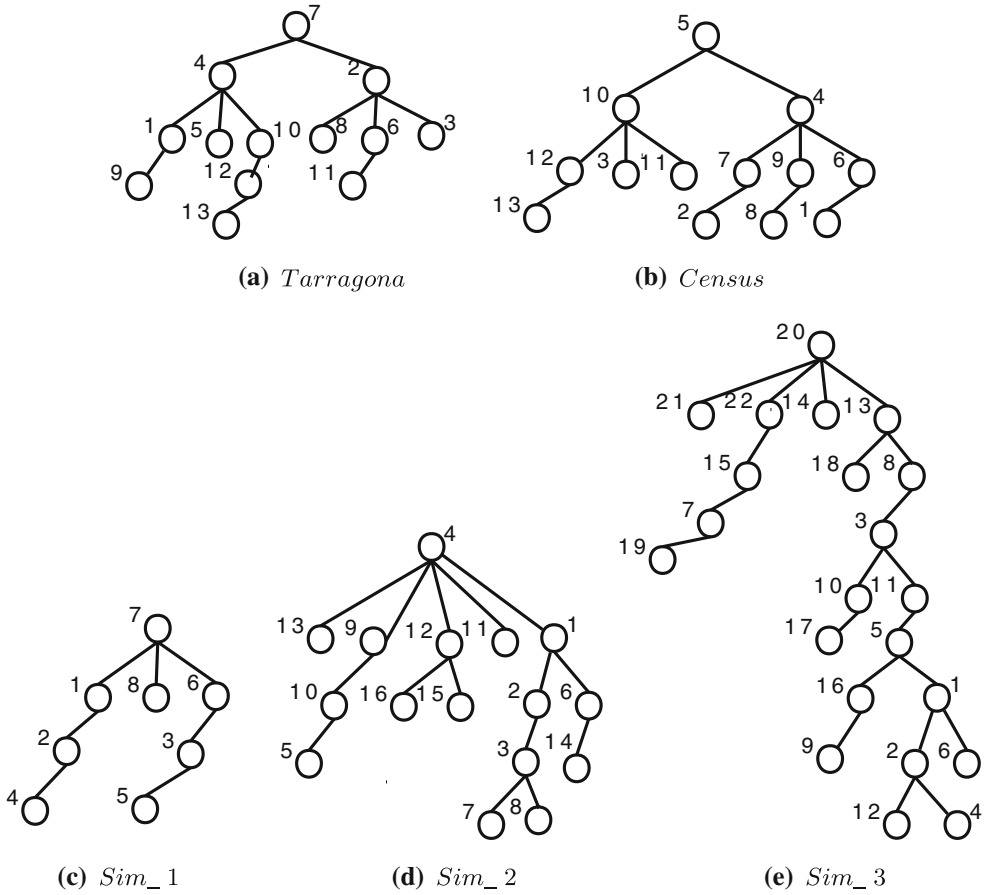


Fig. 17 The best dependence tree for the real and simulated data sets assuming normality

(a) $Tarragona$ **(b)** $Census$ **(c)** Sim_1 **(d)** Sim_2 **(e)** Sim_3

after projecting the entire data set into three variables although the search space was greater than the search space that resulted from using the χ^2 or the EMIM metrics. The minimum value of the IL was equal to 23.10% for Tarragona data set, 16.34% for Census data set, 37.8% for *Sim*1, 51.96% for *Sim*2, and 55.64% for *Sim*3.

Finally, we conclude by stating that each method of calculating the edges weights has its own advantages and disadvantages. We believe that, in practice, the user is the only one who is capable of deciding which is the most suitable metric for the specific data sets. Table 9 summarizes the characteristics of each metric in calculating the edge weights of the graph.

5 Conclusions

In this paper, we have shown how the information about the structure of the dependence between the variables in the micro-data file can be used as a fundamental indicator before invoking any MAT. By using this information, we have proposed a new automated scheme as a pre-processing phase to determine the number and the identity of the variables that are to be used to micro-aggregate the micro-data file. This is achieved by constructing a connected undirected graph whose nodes represent the random variables in the micro-data file, edges represent the statistically dependencies, and the edges weights are computed either

by using the EMIM, χ^2 or the correlation values. The experimental results show that such a methodology involving projecting the multi-variate data sets reducesthe solution space, which further directly reduces the computation time required to search the entire space combinatorially. In spite of this, this methodology leads to a solution whose IL values areclose to the minimum value of the IL that can be obtained by exhaustively searching over the entire search space. The use of these methods for other problems including *k*-anonymity would be an avenue for future research.

References

1. Adam N, Wortmann J (1989) Security-control methods for statistical databases: a comparative study. ACM Comput Surv 21(4):515–556
2. Chow C, Liu C (1968) Approximating discrete probability distributions with dependence trees. IEEE Trans Inf Theory 14(11):462–467
3. Cormen T, Leiserson C, Rivest R (1990) Introduction to algorithms. MIT Press/McGraw-Hill, Cambridge
4. Crises G (2004) Microaggregation for privacy protection in statistical databases. Technical report
5. Cuppen M (2000) Secure data perturbation in statistical disclosure control. PhD thesis, Statistics Netherlands
6. Defays D (1997) Protecting microdata by microaggregation: the experience in Eurostat. Questiio 21:221–231
7. Defays D, Anwar M (1998) Masking micro-data using micro-aggregation. J Off Stat 14(4):449–461
8. Defays D, Anwar N (1995) Micro-aggregation: a generic method. In: Proceedings of the 2nd international symposium on statistical confidentiality. Office for Official Publications of the European Communities, Luxembourg, pp 69–78
9. Defays D, Nanopoulos P (1993) Panels of enterprises and confidentiality: the small aggregates method. In: Proceedings of 92 symposium on design and analysis of longitudinal surveys. Statistics Canada, Ottawa, pp 195–204
10. Domingo-Ferrer J, Mateo-Sanz J (2002) Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans Knowl Data Eng 14(1):189–201
11. Domingo-Ferrer J, Mateo-Sanz J, Oganian A, Torra V, Torres A (2002) On the security of microaggregation with individual ranking: analytical attacks. Int J Uncertain Fuzziness Knowl Based Syst 10(5):477–491
12. Domingo-Ferrer J, Torra V (2002) A quantitative comparison of disclosure control methods for microdata. In: Doyle P, Lane J, Theeuwes J, Zayatz L (eds) Confidentiality, disclosure and data access: theory and practical applications for statistical agencies. North-Holland/Springer, Amesterdam/Berlin, pp 113–134
13. Domingo-Ferrer J, Torra V (2002) Aggregation techniques for statistical confidentiality. In: Aggregation operators: new trends and applications. Physica-Verlag GmbH, Heidelberg, pp 260–271
14. Domingo-Ferrer J, Torra V (2005) Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Min Knowl Disc 11(2):195–212
15. Duda R, Hart P, Stork D (2000) Pattern classification. Wiley-Interscience, New York
16. Fayyoumi E, Oommen B (2006) A fixed structure learning automaton micro-aggregation technique for secure statistical databases. In: Privacy statistical databases, Rome, Italy, pp 114–128
17. Fayyoumi E, Oommen B (2006) On optimizing the *k*-ward micro-aggregation technique for secure statistical databases. In: 11th Austratasian conference on information security and privacy proceeding, Melbourne, Australia, pages 324–335
18. Hansen S, Mukherjee S (2003) A polynomial algorithm for univariate optimal microaggregation. IEEE Trans Knowl Data Eng 15(4):1043–1044
19. Hundepool A, Wetering A, Ramaswamy R, Franconi L, Capobianchi A, Wolf P, Domingo-Ferrer J, Torra V, Brand R, Giessing S (2004) M-ARGUS Version 4.0 Software and User's Manual
20. Kim J, Winkler W (1995) Masking microdata files. In: Proceedings of the section on survey research methods, pp 114–119
21. Laszlo M, Mukherjee S (2005) Minimum spanning tree partitioning algorithm for microaggregation. IEEE Trans Knowl Data Eng 17(7):902–911
22. Li Y, Zhu S, Wang L, Jajodia S (2002) A privacy-enhanced microaggregation method. In: FoIKS'02: proceedings of the second international symposium on foundations of information and knowledge systems. Springer, London, pp 148–159
23. Mas M (2006) Statistical data protection techniques. Technical report, Eustat: Euskal Estatistika Erakundea, Instituto Vasco De Estadistica
24. Mateo-Sanz J, Domingo-Ferrer J (1998) A comparative study of microaggregation methods. Questiio 22(3):511–526
25. Mateo-Sanz J, Domingo-Ferrer J (1999) A method for data-oriented multivariate microaggregation. In: Proceedings of statistical data protection'98. Office for Official Publications of the European Communities, Luxembourg, pp 89–99
26. Nin J, Herranz J, Torra V (2008) How to group attributes in multivariate microaggregation. Int J Fuzziness Knowl Based Syst 16(1):121–138
27. Oganian A, Domingo-Ferrer J (2001) On the complexity of optimal microaggregation for statistical disclosure control. Stat J United Nations Econ Comm Europe 18(4):345–354
28. Oommen B, Fayyoumi E (2007) A novel method for micro-aggregation in secure statistical databases using association and interaction. In: Information and communications security, 9th international conference on information and communications security. LNCS 4861, Springer, Berlin, pp 126–140
29. Panaretos J, Tzyvidis N (2001) Aspects of estimation procedures at Eurostat with some emphasis on over-space harmonisation. In: HERCMA 2001 conference
30. Sanchez J, Urrutia J, Ripoll E (2004) Trade-off between disclosure risk and information loss using multivariate microaggregation: a case study on business data. In: Domingo-Ferrer J, Torra V (eds) Privacy in statistical databases: CASC project international workshop, PSD 2004 proceedings, Barcelona, Spain. Springer, Berlin, pp 307–322
31. Solanas A, Martíínez-Ballestéé A (2006) V-MDAV: a multivariate microaggregation with variable group size. In: 17th COMPSTAT symposium of the IASC, Rome
32. Torra V (2004) Microaggregation for categorical variables: a median based approach. In: Domingo-Ferrer J, Torra V (eds) Privacy in statistical databases: CASC project international workshop, PSD 2004 proceedings, Barcelona, Spain. Springer, Berlin, pp 162–174
33. Valiveti R, Oommen B (1992) On using the chi-squared metric for determining stochastic dependence. Pattern Recogn Lett 25(11):1389–1400
34. Valiveti R, Oommen B (1993) Determining stochastic dependence for normally distributed vectors using the chi-squared metric. Pattern Recogn Lett 26(6):975–987